Resampling-Based Multiple Hypothesis Testing Procedures for Genetic Case-Control Association Studies

Bingshu E. Chen,[1]* Lori C. Sakoda,[2] Ann W. Hsing,[2] and Philip S. Rosenberg[1]

[1] *Biostatistics Branch, Department of Health and Human Services, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Rockville, Maryland*

[2] *Hormonal and Reproductive Epidemiology Branch, Department of Health and Human Services, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Rockville, Maryland*

In case-control studies of unrelated subjects, gene-based hypothesis tests consider whether any tested feature in a candidate gene—single nucleotide polymorphisms (SNPs), haplotypes, or both—are associated with disease. Standard statistical tests are available that control the false-positive rate at the nominal level over all polymorphisms considered. However, more powerful tests can be constructed that use permutation resampling to account for correlations between polymorphisms and test statistics. A key question is whether the gain in power is large enough to justify the computational burden. We compared the computationally simple Simes Global Test to the **min P** test, which considers the permutation distribution of the minimum p-value from marginal tests of each SNP. In simulation studies incorporating empirical haplotype structures in 15 genes, the **min P** test controlled the type I error, and was modestly more powerful than the Simes test, by 2.1 percentage points on average. When disease susceptibility was conferred by a haplotype, the **min P** test sometimes, but not always, under-performed haplotype analysis. A resampling-based omnibus test combining the **min P** and haplotype frequency test controlled the type I error, and closely tracked the more powerful of the two component tests. This test achieved consistent gains in power (5.7 percentage points on average), compared to a simple Bonferroni test of Simes and haplotype analysis. Using data from the Shanghai Biliary Tract Cancer Study, the advantages of the newly proposed omnibus test were apparent in a population-based study of bile duct cancer and polymorphisms in the prostaglandin-endoperoxide synthase 2 (*PTGS2)* gene. *Genet. Epidemiol.* 2006. Published 2006 Wiley-Liss, Inc.[†]

Key words: case-control studies; haplotypes; single nucleotide polymorphsim; permutation test; biliary tract neoplasms; prostaglandin-endoperoxide synthase 2

Contract grant sponsor: Intramural Research Program of the NIH, National Cancer Institute, Division of Cancer Epidemiology and Genetics.

*Correspondence to: Bingshu E. Chen, PhD, Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, 6120 Executive Blvd, Executive Plaza South, Room 8089, Rockville, MD 20852-7244. E-mail: cheneric@mail.nih.gov

Received 11 November 2005; Accepted 1 May 2006
Published online in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20162

INTRODUCTION

The genetic case-control study is a leading method to identify candidate genes associated with the risk of complex diseases [Schork, 2002]. This design is particularly popular for diseases with a comparatively late age-at-onset, including many cancers. A single study can probe a large number of single nucleotide polymorphisms (SNPs), and the resulting multiple comparisons problem has been widely recognized [Schork et al., 2000; Emahazion et al., 2001]. To date, the accumulating literature has demonstrated a high prevalence of false-positive reports [Ioannidis et al., 2001]. Meta-analysis provides one approach to demonstrate replication validity [Ioannidis et al., 2002]. For certain diseases, such as breast and prostate cancers, consortia have been established to coordinate analyses and pool results, thereby creating huge datasets for hypothesis testing, replication, and estimation (http://epi.grants.cancer.gov/BPC3/cohorts.html). For associations that can be studied by meta-analysis or consortia, these approaches might demonstrate replication validity, and thereby resolve the multiple comparisons issue. However, for less common diseases, and for common or uncommon diseases arising in special populations, it may not be practical to assemble such large studies. Therefore, especially in these situations, as in general, a challenge for investigators is to honestly assess the statistical significance of a candidate gene association hypothesized *a priori*, in the context of a single study, using the most powerful available statistical test that maintains control of the overall gene-wide false-positive rate.

An example of such a study is provided by the Shanghai Biliary Tract Cancer Study (SBTCS) [Sakoda et al., 2006]. The SBTCS is investigating the etiology of biliary tract cancer among residents of Shanghai China, a region where these comparatively rare tumors are increasing in incidence [Hsing et al., 1998]. Using a population-based case-control design, the SBTCS study enrolled more than 95% of all incident cases arising over the period from June 1997 through May 2001. The study enrolled 627 biliary tract cancer cases and 959 age- and sex-matched control subjects. Of them, 411 cases and 786 controls provided adequate blood samples for genetic testing. Among the 411 cases were 237 gallbladder, 127 bile duct, and 47 ampulla of Vater cancers; genetic associations with each type of biliary tract cancer are of interest. However, with the available numbers of cases, efforts to identify candidate genes with moderate effects will encounter issues of power and sample size. Therefore, it is desirable to use the most powerful available statistical tests.

Several analytical approaches are available to test the hypothesis that any variant in a candidate gene is associated with disease. In addition to haplotype analysis, other gene-wide tests include SNP-based tests, and omnibus tests [Rosenberg et al., 2006] that combine SNP and haplotype analysis. Standard tests can be conducted using output from existing computer programs. However, resampling-based tests that account for correlation between test statistics should control the false-positive rate at the nominal level and have higher power. A key question is whether the gain in power is large enough to justify the computational burden.

In this report, we study the performance of some resampling-based tests in simulations conducted over two panels of candidate genes. One panel includes nine candidate genes implicated in diabetes and autoimmune disorders, and haplotypes observed in a population of European ancestry. The other panel includes six candidate genes in the chronic inflammation pathway, and haplotypes observed in an Asian population (the SBTCS). As we show, the proposed methods offer gains in power that appear large enough to be useful in practice. We illustrate the testing procedures using data from the SBTCS. A technical appendix presents computational details.

METHODS

Each of the statistical tests proposed here are "gene-based" association tests [Neale and Sham, 2004] that consider whether any tested "features" in a candidate gene—SNPs, haplotypes, or both—are associated with disease. Gene-based tests are designed to control the family-wise type I error rate for the complete null hypothesis that no tested feature of a candidate gene is associated with disease.

TEST PROCEDURES

An investigator with SNP data in hand can use a number of analytical approaches, including the following:

1. Test each SNP variant marginally without any adjustment for multiple comparisons.
2. Test each SNP variant marginally, adjusting for multiple comparisons using the Simes Global Test [Simes, 1986], which provides a more powerful alternative to a Bonferroni test.
3. Test each SNP variant marginally, adjusting for multiple comparisons with the Bonferroni test using the number of independent tests obtained from the spectral decomposition of matrices describing the pair-wise linkage disequilibrium (LD) between SNPs [Nyholt, 2004; Nicodemus et al., 2005], or the number of LD blocks. The later method of adjustment may not control the type I error rate under high-LD scenarios [Nicodemus et al., 2005] and is not considered further here.
4. Perform a haplotype analysis to determine whether the haplotype frequency distribution differs between cases and controls, accounting for uncertainty about the linkage phase.
5. Combine approaches 2 and 4, using a Bonferroni correction to correct for the fact that two tests were conducted; twice the smaller of the two raw p-values is the test statistic [Rosenberg et al., 2006].

Approach 1 does not control the type I error rate and is highly prone to false positives [Nicodemus et al., 2005; Rosenberg et al., 2006]. Approaches 2–5 do control the type I error rates, and therefore provide valid gene-based tests. Approaches 1–3 require standard output from logistic regression [Prentice and Pyke, 1979]. Given M SNPs in a candidate gene, one computes M marginal likelihood ratio tests. For example, one might use a trend test [Armitage, 1955] that considers the number of copies of a variant SNP. For Approach 4, several haplotype frequency tests (HFTs) have been proposed [Zhao et al., 2000; Fallin et al., 2001; Schaid et al., 2002; Epstein and Satten, 2003] and computation tools for haplotype analysis are available, including *SAS PROC HAPLOTYPE* [SAS Institute Inc., 2002], *HAPLO.SCORE* [R Development Core Team, 2004], *PHASE* [Stephens et al., 2001], and the software provided by Satten [Epstein and Satten, 2003]. HFTs can be adjusted for covariates [Li et al., 2003; Zhao et al., 2003; Lake et al., 2003]. Approach 5 requires only the two summary p-values obtained for Analyses 2 and 4 and simple arithmetic; therefore, we call it the "simple" omnibus test.

The Simes Global Test (Approach 2) has comparatively high power when disease susceptibility is conferred by a SNP, but it may have a substantial false-negative rate if disease susceptibility is conferred by a haplotype [Rosenberg et al., 2006]. Conversely, haplotype analysis (Approach 4) has comparatively high power when disease susceptibility is conferred by a haplotype, but it may have a substantial false-negative rate if disease susceptibility is conferred by an SNP. The simple omnibus test (Approach 5) tracks the more powerful of the SNP-based or haplotype-based analysis, which is generally unknown. In the next section, we show how to extend Approaches 2, 4, and 5 so that each continues to control the type I error rate but has higher power. Specifically, the Simes Global Test may not fully exploit the extent of LD between SNPs; some applications of haplotype analysis may not take advantage of the possibility that there is only limited haplotype diversity; the simple omnibus test does not account for correlation between the base tests. The tests described below—the **min P** test [Westfall and Young, 1993; Westfall et al., 2002], the "directed" HFT, and the resampling-based omnibus tests—overcome these respective limitations.

THE min P TEST

We evaluate whether the **min P** test provides more statistics power than the Simes Global Test (Approach 2), and if so, by how much. The **min P** test was first suggested by Westfall et al. [Westfall and Young, 1993; Westfall et al., 2002]. Both the **min P** test and the Simes Global Test consider p-values for a set of SNP-disease associations one-at-a-time, marginally over all other SNPs. With the **min P** test, inference is based on the permutation distribution of the minimum of the ordered p-values, which takes the correlations into account. In contrast, the Simes Global Test uses a non-iterative procedure to adjust the minimum observed p-value for multiplicity. The maximum χ^2 test [de Bakker et al., 2005], which compares the maximum of the marginal χ^2 test statistics from each SNP with the distribution of such statistics under null hypothesis using permutation, is closely related to the **min P** test described here.

Suppose there are M SNPs in a candidate gene, and the marginal test for the jth SNP yields an observed p-value p_j. We denote the **min P** test statistic as

$$WYZ = \min_{1 \le j \le M} p_j.$$

Under the complete null hypothesis, the case-control indicators can be permuted B times to generate a set of permutation samples. Let p^*_{jb} be the p-value for the jth SNP in the bth permutation sample, obtained by shuffling case-control indicators. The permuted **min P** statistic is given by

$$WYZ^*_b = \min_{1 \le j \le M} p^*_{jb}$$

and the permutation-based p-value for the **min P** test WYZ, called p^{WYZ}, is the proportion of $\{WYZ^*_b\}^B_{b=1}$ that are equal to or smaller than the observed **min P** statistic WYZ [Westfall and Young, 1993]. At least $B = 1{,}000$ permutated datasets are needed to obtain a reasonably accurate estimate of p^{WYZ}. Missing SNP genotype data are accommodated; each marginal test makes use of all subjects with available data for that SNP.

DIRECTED AND GLOBAL HFTS

Current laboratory methods do not permit large-scale resolution of the gametic phase in studies of unrelated subjects. Therefore, for each individual who is heterozygous for more than one SNP in a candidate gene, the number of copies of each variant SNP allele is known, 0, 1, or 2, but it is not known which SNP alleles are present on each chromosome. Under the assumption of Hardy-Weinberg Equilibrium (HWE), haplotype frequencies can be estimated from unphased SNP genotype data by maximum likelihood using the

EM algorithm [Dempster et al., 1977; Excoffier and Slatkin, 1995]; missing SNP genotype data can be accounted for in the likelihood calculations.

An HFT compares the reconstructed haplotype frequency distributions in cases versus controls. A global HFT, HFT_G, compares the frequencies of all haplotypes that are inferred to be present in cases or controls and tests for significance using a permutation test [Fallin et al., 2001]. In contrast, a "directed" HFT considers frequencies only for comparatively common haplotypes, say, those with an estimated frequency of 5% or higher in controls. All the less common haplotypes are pooled into an "other" category. Because the frequencies of the common haplotypes can be estimated stably, we construct a directed test, HFT_D, using a Wald-type test statistic and closed-form expressions for the variance-covariance matrix of the case-control differences in haplotype frequencies (Appendix). Here, we evaluate whether a directed test provides more statistical power than a global test in selected scenarios, and if so, by how much. As a benchmark in simulation studies, we also evaluate HFT_I, an ideal HFT that contrasts the frequencies of the common haplotypes using phase-known data.

RESAMPLING-BASED OMNIBUS TESTS

We attempt to make the simple omnibus test (Approach 5) more powerful by using base tests with potentially greater power, and by accounting for correlation between the base tests. As with the simple omnibus test, we combine a SNP-based test and a haplotype-based test. For the SNP-based test, we replace the Simes Global Test by the **min P** test statistic, and for the haplotype-based test, we consider both the directed and the global HFT. Therefore, we proposed a version of the omnibus test statistic with

$$OMNI = \min\left(p^{HFT}, p^{WZY}\right)$$

where p^{HFT} is the p-value of HFT_D or HFT_G, and p^{WZY} is the p-value of the **min P** statistic. The distribution of this omnibus test statistic can be estimated from its permutation distribution. For the bth permuted dataset, one can conduct both the SNP-based test and the haplotype-based test, yielding p-values p_b^{WZY*} and p_b^{HFT*}. Here p_b^{WZY*} is the proportion of $\left\{WYZ_{b'}^*\right\}_{b'=1}^B$ that are equal to or smaller than WYZ_b^*. For the directed HFT, $p_b^{HFT_D{}^*}$ can be obtained from a central χ^2 distribution. For the global HFT, $p_b^{HFT_G{}^*}$is the proportion of $\left\{HFT_{G,b'}{}^*\right\}_{b'=1}^B$ that are equal to or greater than $HFT_{G,b}{}^*$. The bth permuted value of the resampling-based omnibus test statistic is

$$OMNI_b^* = \min\left(p_b^{HFT*}, p_b^{WZY*}\right).$$

The p-value for the omnibus test $OMNI$, called p^{OMNI}, can be estimated by the proportion of $\left\{OMNI_b^*\right\}_{b=1}^B$ that is equal to or smaller than the test statistics $OMNI$. Note that a single set of permutation samples is used to compute each test, similar to an algorithm used for micro-array studies [Ge et al., 2003].

CANDIDATE GENE PANELS AND SIMULATION STUDIES

To investigate the performance of the proposed methods, we considered two panels of candidate genes (Table I). The first panel (the "Johnson Panel") includes *CASP8*, *CASP10*, *CFLAR*, *CTLA4*, *GAD2*, *H19*, *INS*, *SDF1*, and *TCF8* (122 SNPs were genotyped, with 59 common SNPs in nine genes) in 135 kb of DNA [Johnson et al., 2001]. In the Johnson Panel, on average, there are 6.6 common SNPs (with frequency ≥5%) and 4.9 common haplotypes (with frequency ≥5%) per gene.

The second panel (the "SBTCS Panel") includes *IL10*, *IL1A*, *IL1B*, *IL4*, *PTGS2*, and *TNF*; these six genes are members of the chronic inflammation pathway. The SBTCS Panel was constructed as follows. For each gene, 6, 3, 3, 5, 8, and 5 SNPs respectively from the SNP500Cancer database (http://snp500cancer.nci.nih.gov) were selected for genotyping in 44 kb of DNA. Genomic DNA was extracted from buffy coat samples of cases and controls. SNPs were genotyped using TaqMan assays at the Core Genotyping Facility (http://cgf.nci.nih.gov) of the National Cancer Institute (Rockville, MD, USA). The numbers of genotyped SNPs in each candidate gene with any variation in cases or controls were 3, 2, 3, 5, 5, and 5, respectively. In these 23 SNPs with variation, the mean fraction of missing genotype data was 1.2% (range: 0.3–3.6%). Genotype frequencies in the population controls were in HWE. Twenty-one of these 23 SNPs had a minor allele frequency ≥5%. The mean number of common SNPs per gene was 3.5, and the mean number of common haplotypes per gene was 2.5. Haplotypes and their corresponding frequencies are shown in Figure 1. Using genotype data from controls, the consistent haplotypes were enumerated, and the corresponding haplotype frequencies in the SBTCS population were estimated using the EM algorithm.

TABLE I. Type I error rates under the complete null hypothesis

GENE	Nyholt	**min P**	Global HFT	Directed HFT[a]	Omnibus test (**min P**+Global HFT)	Omnibus test (**min P** +Directed HFT[a])
Johnson Panel						
CASP8	0.042	0.051	0.045	0.046	0.054	0.045
CASP10	0.039	0.055	0.044	0.039	0.041	0.053
CFLAR	0.050	0.050	0.046	0.036	0.046	0.049
CTLA4	0.054	0.050	0.032	0.033	0.038	0.055
GAD2	0.039	0.054	0.052	0.055	0.058	0.048
H19	0.041	0.054	0.055	0.039	0.054	0.043
INS	0.042	0.051	0.052	0.038	0.053	0.042
SDF1	0.036	0.051	0.050	0.040	0.048	0.042
TCF8	0.044	0.049	0.052	0.045	0.044	0.039
SBTCS Panel						
IL10	0.055	0.046	0.049	0.048	0.051	0.049
IL1A	0.054	0.051	0.052	0.047	0.055	0.048
IL1B	0.042	0.048	0.053	0.042	0.045	0.044
IL4	0.056	0.054	0.046	0.053	0.048	0.047
PTGS2	0.048	0.043	0.055	0.040	0.052	0.042
TNF	0.059	0.047	0.046	0.049	0.044	0.051

Type I error rates were estimated from $B = 1{,}000$ replications of studies with $n_1 = n_0 = 300$ cases and controls under the null hypothesis. The nominal α level for each procedure was 0.05.

[a]For the Directed HFT, haplotypes in the population with frequency greater than 5% were used to define the Wald test.

In simulation studies for each gene in both panels, cohorts of 100,000 individuals were generated with haplotypes randomly assigned assuming HWE. Then for each individual, disease status was assigned according a Bernoulli random number with probability

$$P(Disease = 1|x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

where x equals the number of copies of the disease susceptibility SNP or haplotype, and β_1 is the logarithm of the relative risk. The baseline parameter β_0 was 0.01. Finally, case-control samples were obtained by randomly selecting $n_0 = 300$ controls and $n_1 = 300$ cases from each cohort [Rosenberg et al., 2006]. Replicate datasets were created, both under the null, and with each common SNP and haplotype (with frequency greater than 5%) in turn associated with disease according to a codominant logistic model with a relative risk of 1.5 per copy of each associated SNP or haplotype. Type I error rates and power were computed from 1,000 replicate studies. For the resampling-based tests, 1,000 permutation samples per replicate study were used to estimate the empirical distribution function of the test statistics under the null.

RESULTS

SIMULATION RESULTS

The proposed gene-based tests have the correct type I error rate (Table I): for each gene in both panels, the percentage of false positives is close to the nominal level of 0.05. In particular, we observed that Nyholt's approach controls the Type I error rate at the nominal level. In contrast, despite the extensive LD between SNPs, the type I error rate is very high if each SNP is tested without any adjustment for multiplicity. The percentage of false positives (computed by simulating under the complete null) ranged from 8% to 23% over genes in the SBTCS Panel (with 3.5 common SNPs/gene on average), and from 18% to 45% in the Johnson Panel (with 6.8 common SNPs/gene on average (data not shown)).

The **min P** test has higher power than the Simes Global Test (Table II). Table II shows the mean power for the **min P** test, and the Simes Global Test, evaluated gene-by-gene in scenarios where each common SNP and haplotype (frequency $>5\%$) in turn was associated with disease. For example, in *CASP8*, the power of the **min P** test and the Simes Global Test was evaluated for seven common SNPs and five common haplotypes.

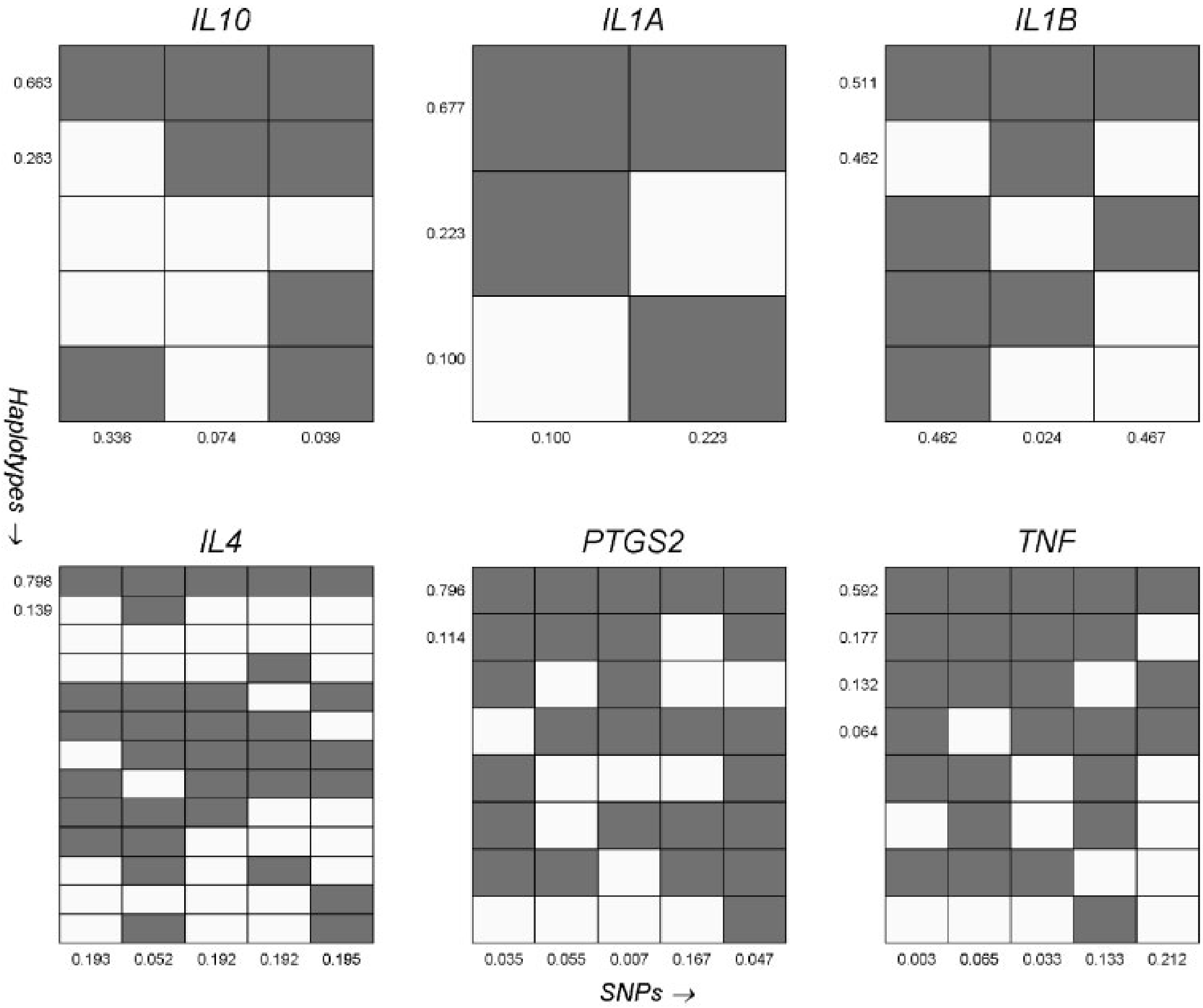

Fig. 1. Haplotype structures in the SBTCS Panel. Panels correspond to genes, rows to haplotypes, and columns to SNPs. The *y*-axis labels show common haplotype frequencies. Rare haplotypes (un-labeled) are pooled in a haplotype analysis. The *x*-axis labels show SNP allele frequencies.

On average over these 12 scenarios, the **min P** test had 3.0 percentage points higher power (61.2 versus 58.2 percent power, for **min P** and Simes Global Test, respectively). Averaged over genes, the power of the **min P** test versus the Simes Global Test was 2.2 percentage points higher in the Johnson Panel and 2.0 percentage points higher in the SBTCS Panel. We also observed that the power of Nyholt's approach lies in between Simes Global Test and the **min P** test (data not shown).

For haplotype analysis, HFT_D had higher power than HFT_G, when disease susceptibility was conferred by a haplotype with frequency ≥5% (Fig. 2; Johnson Panel with disease susceptibility conferred by each haplotype). Indeed, HFT_D was nearly as powerful as the ideal test HFT_I. Compared to HFT_I, on average in Figure 2, HFT_D had 3.8 percentage points less power. This gap in power reflects the impact of phase ambiguity on analyses restricted to the common haplotypes.

Comparing HFT_D to HFT_G (both tests based on unphased data), on average, the directed test had 6.6 percentage points higher power in the Johnson Panel, and 4.7 percentage points higher power in the SBTCS Panel (Table II). The gap in power between the global and directed tests reflects the additional degrees of freedom that are incorporated into the global test. The directed and global tests had similar power for the *CFLAR* gene; three of the four haplotypes in this gene have a frequency ≥5%, so little difference is expected between the two tests.

The resampling-based omnibus tests dominated the simple omnibus test, regardless of whether

TABLE II. Summary of average power for min P, HFT_D, and resampling-based omnibus tests in studies with 300 cases and controls

	Average power (percentage points)					
	SNP Tests		Haplotype frequency tests		Omnibus tests[a]	
	min P	Simes	HFT_D	HFT_G	Resampling	Simple
Johnson Panel[b]						
CASP8	61.2	58.2	50.1	37.6	52.3	47.6
CASP10	61.7	60.4	53.8	50.2	58.1	52.8
CFLAR	73.3	73.2	69.3	69.9	72.4	67.5
CTLA4	52.5	51.2	34.5	31.8	40.0	34.1
GAD2	58.6	56.8	59.5	55.4	63.5	58.1
H19	71.4	66.4	49.7	33.8	55.5	50.1
INS	61.3	58.8	58.8	46.9	62.0	55.7
SDF1	57.0	56.9	43.4	39.0	50.9	44.2
TCF8	49.9	47.6	33.7	28.3	38.7	33.8
SBTCS Panel						
IL10	61.7	60.1	60.9	58.4	63.1	56.9
IL1A	68.5	67.4	69.0	68.2	69.4	62.6
IL1B	89.4	89.9	90.4	82.3	90.0	86.7
IL4	62.0	58.6	54.5	42.3	61.8	54.7
PTGS2	38.3	34.7	37.0	32.9	41.1	35.8
TNF	46.8	43.4	44.8	44.0	48.7	42.4

[a]The resampling omnibus tests combine **min P** and HFT_D, and the simple omnibus test is the Bonferroni correction applied to Simes Global Test and HFT_D.

[b]In the Johnson Panel, the haplotype frequency tests HFT_D and HFT_G, and the resampling and simple omnibus tests were compared in scenarios where each common haplotype in turn conferred susceptibility ($RR = 1.5$ per copy); all other comparisons were evaluated over scenarios where each common SNP and haplotype (frequency >5%) in turn conferred susceptibility.

disease susceptibility was conferred by an SNP or by a haplotype (Fig. 3; SBTCS Panel). Here, the resampling-based omnibus test combines **min P** and HFT_D, and the simple omnibus test combines the Simes Global Test and HFT_D using a Bonferroni correction. On average, the resampling-based omnibus test had 5.8 percentage points higher power in the SBTCS Panel and 5.6 percentage points higher power in the Johnson Panel (Table II).

Overall, the omnibus test combining **min P** and HFT_D provided higher statistical power for tests of the "common-disease, common-variant" hypothesis than the omnibus test combining **min P** and HFT_G. For example, in the Johnson Panel with disease susceptibility conferred by each haplotype, on average, the omnibus test incorporating HFT_D had 5.2 percentage points higher power than the omnibus test incorporating HFT_G, and the maximum gain in power was 27 percentage points for the gene *H19* which has 15 total haplotypes but only five common ones.

Figures 4 and 5 show power curves for the omnibus test combining **min P** and HFT_D, applied to the two gene panels. These data show that haplotype analysis is not necessarily the most powerful approach to detect association conferred by a common haplotype. In every gene in the Johnson Panel (Fig. 4), at least one haplotype exists for which the **min P** test was more powerful. To a lesser extent, this phenomenon was also observed in the SBTCS Panel (Fig. 5, *IL4* and *TNF*). Conversely, the **min P** test was always more powerful when disease susceptibility was conferred by a common SNP (data not shown). In all scenarios considered, the resampling-based omnibus test closely tracked the more powerful of the two component tests. For other choices of relative risks (for example, $RR = 1.25$ and 2.0), we observed similar patterns of power gains to the $RR = 1.5$ scenarios (data not shown).

APPLICATION TO THE SBTCS

In the SBTCS, the prostaglandin-endoperoxide synthase 2 (*PTGS2*) gene is of particular interest. *PTGS2* encodes for one isoform of the enzyme PTGS, also commonly known as cycloxygenase (COX), which converts arachidonic acid into prostaglandins. Given that PTGS2 is induced by

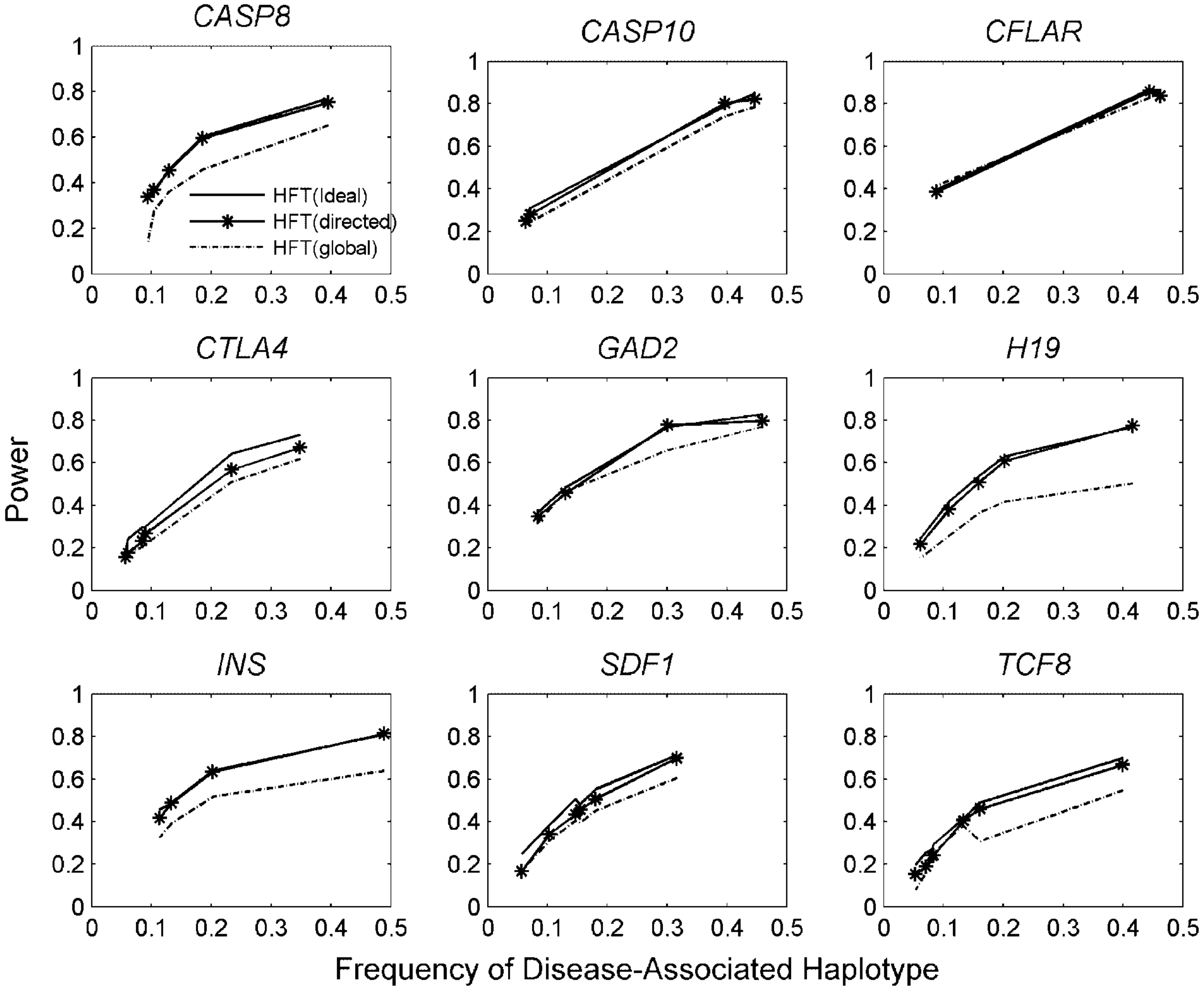

Fig. 2. Power curves of haplotype frequency tests for the Johnson Panel when disease susceptibility is conferred by a haplotype with a codominant effect. Panels correspond to genes, abscissas to susceptibility haplotype (arranged by frequency), and ordinates to power. Values of the power were determined from $B = 1{,}000$ replications of studies with $n_1 = n_0 = 300$ cases and controls. Power curves for ideal phase-known HFT, directed HFT, and global HFT are shown.

proinflammatory and mitogenic factors and overexpressed in malignant biliary tissue, it was hypothesized that variants in the *PTGS2* gene may alter the expression or function of its encoded enzyme, thereby modulating inflammatory processes that influence cancer susceptibility.

Table III reports marginal SNP-based analysis and haplotype analysis, comparing 127 bile duct cancer cases to 786 population controls. By SNP analysis, the SNP 4: T>C allele showed a significant association with bile duct cancer (nominal p-value of 0.0041), with a relative risk of 1.63 per copy of the C allele (95% confidence interval: 1.17–2.25). The p-value of Simes Global Test was 0.0205, and the p-value of the **min P** test was 0.0145 (based on 10,000 permutations); both of these tests adjust for multiplicity over the SNPs.

To construct a directed HFT, we focused on the common haplotypes with a frequency greater than 1% in the population controls. Four common haplotypes (shown in Table III) were used to construct the directed HFT, which yielded a χ^2 test statistic of 10.96 on 4 degrees of freedom, and p-value of 0.027, indicating a statistically significant difference in the frequencies of the common haplotypes in cases versus controls. The frequency of the most common haplotype (00000) was 79.6% in the population controls, versus 72.0% in the cases (Table III). For haplotypes (00010) and (01110), the estimated frequencies were 11.4% and 4.7% in the population controls and 15.0%

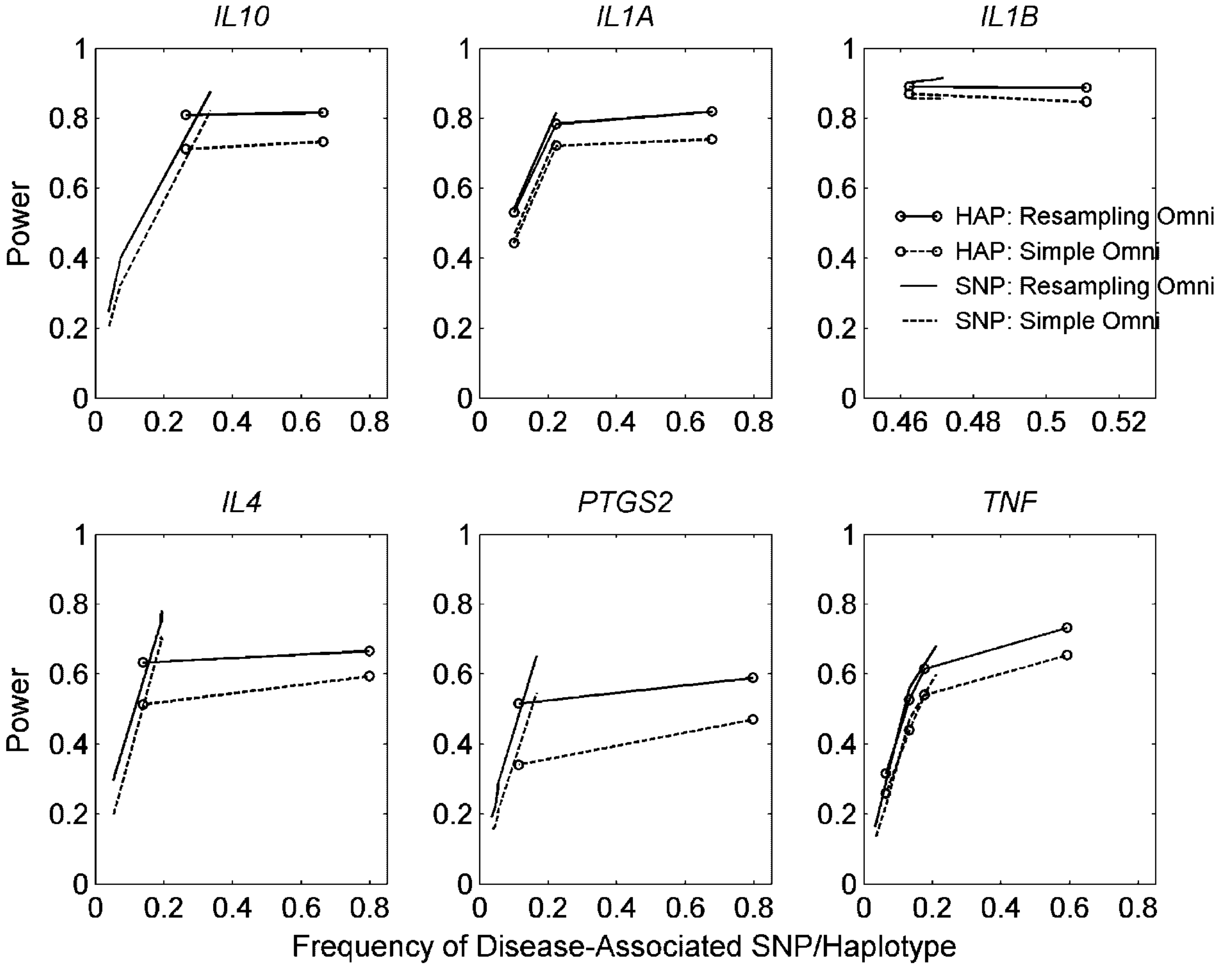

Fig. 3. Power curves of omnibus tests for the SBTCS Panel when disease susceptibility is conferred by an SNP or a haplotype with a codominant effect. The panel design is similar to Figure 2, except that abscissas include scenarios where SNP or haplotypes are associated with disease, and the range of abscissas for *IL1B* gene is from 0.46 to 0.52. Power curves for resampling-based omnibus test and simple omnibus test are shown.

and 7.1% in cases. In contrast, the global HFT considering 11 haplotypes (eight haplotypes shown in Figure 1, and three rare haplotypes inferred in the cases) yielded a non-significant result with a p-value of 0.1236.

If one uses the simple omnibus test to combine the Simes Global Test and the directed HFT, the p-value is 0.041, which equals twice the smaller of 0.0205 (Simes Global Test) and 0.027 (HFT_D). A smaller p-value of 0.015 is obtained using the resampling-based omnibus test that combines $\mathbf{min\,P}$ and HFT_D. This p-value accounts for the fact that five SNPs and four common haplotypes were considered using two analytical approaches. The results indicate that SNP 4 is positively associated with bile duct cancer, as are the haplotypes (00010) and (01110) that contain it.

The difference between the p-values obtained using the simple omnibus test (0.041) and the resampling-based omnibus test (0.015) is important. Consider a planned analysis to test each of the six genes in the SBTCS Panel for association with bile duct cancer, adjusting for multiplicity over the panel using the Benjamini-Hochberg False Discovery Rate (FDR) [Benjamini and Hochberg, 1995]. In the worst-case scenario, the simple omnibus test would yield an FDR-adjusted q-value for *PTGS2* of 0.246 (6×0.041), whereas the resampling-based omnibus test would yield an FDR-adjusted q-value of 0.09 (6×0.015). Hence, *PTGS2* would be "discovered" at the 10% level using the resampling-based omnibus test but not with the simple omnibus test. A similar argument holds if one planned to apply a Bonferroni correction.

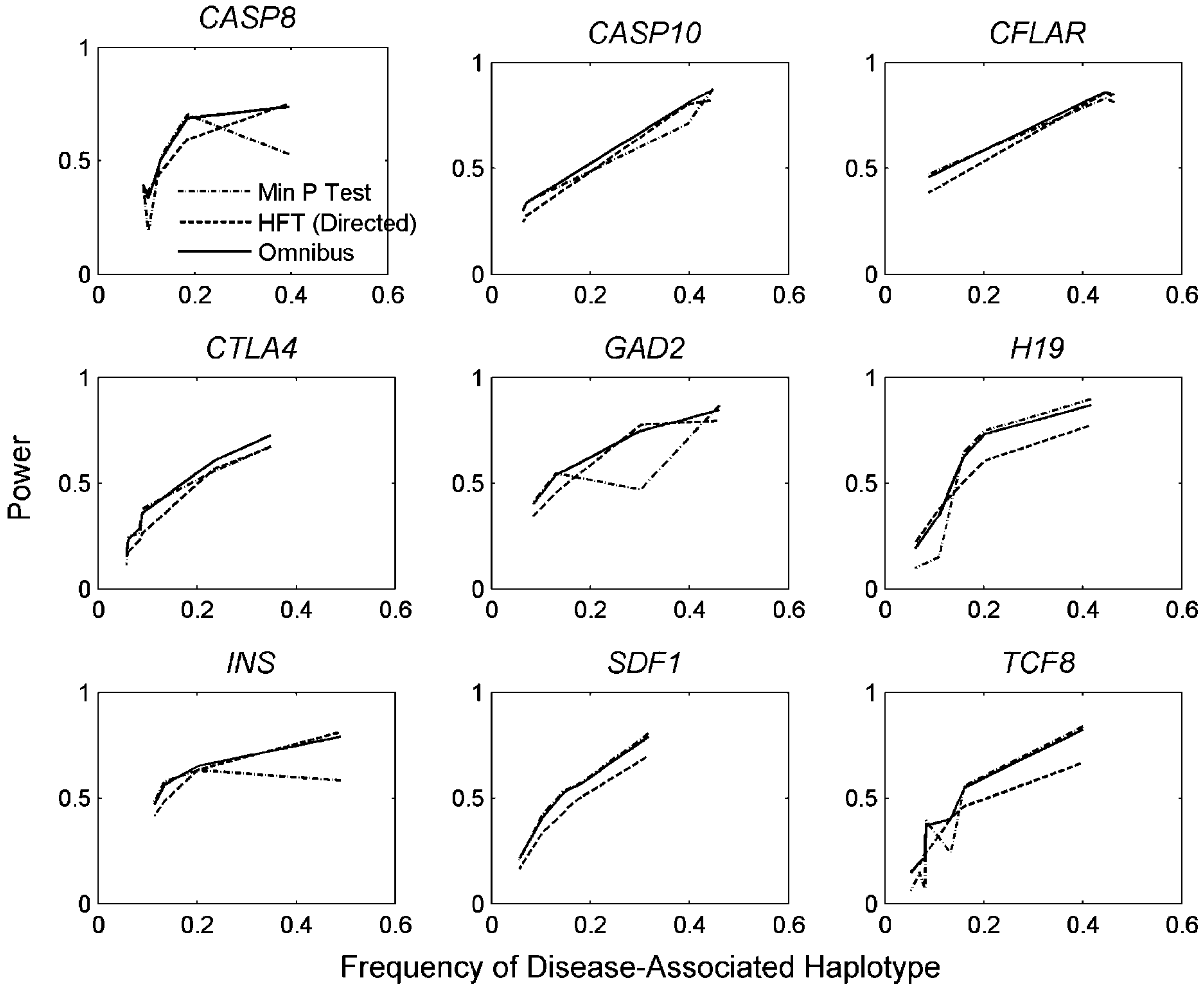

Fig. 4. Power curves of omnibus tests for the Johnson Panel when disease susceptibility is conferred by a haplotype with a co-dominant effect. The panel design is the same as in Figure 2. Power curves for min P test, directed HFT, and resampling-based omnibus test are shown.

DISCUSSION

To avoid false-negative results in a candidate gene analysis, it makes good sense to consider both haplotype analysis and the marginal associations of disease with each SNP. However, the false-positive rate can be very high if multiple SNPs are analyzed without any adjustment for multiple comparisons, ranging from 8% to 45% across the 15 genes considered here. These false-positive rates pertain to analysis of a *single* candidate gene. Of course, the chance of obtaining at least one false-positive gene increases rapidly with the number of candidate genes.

Gene-based test provide a means to avoid such false positives. "Off-the-shelf" methods can be applied using widely available software. However, the resampling-based approaches investigated here offer genuine gains in power (Table II). These gains were achieved in part by accounting for the correlation between test statistics. Overall, we recommend the resampling-based omnibus test that combines the **min P** test with a directed HFT. This test makes use of all available genotype data, accounts for phase ambiguity in the haplotype analysis, adjusts for multiplicity over all SNPs and haplotypes considered using two analytical approaches, and closely tracks the more powerful of the component tests. Of course, in a particular study, one cannot know whether the resampling-based methods are providing a modest or substantial gain in power. However, our results suggest that these new approaches should be useful in practice. The procedures operated similarly in both panels considered, and the SBTCS Panel was derived from widely

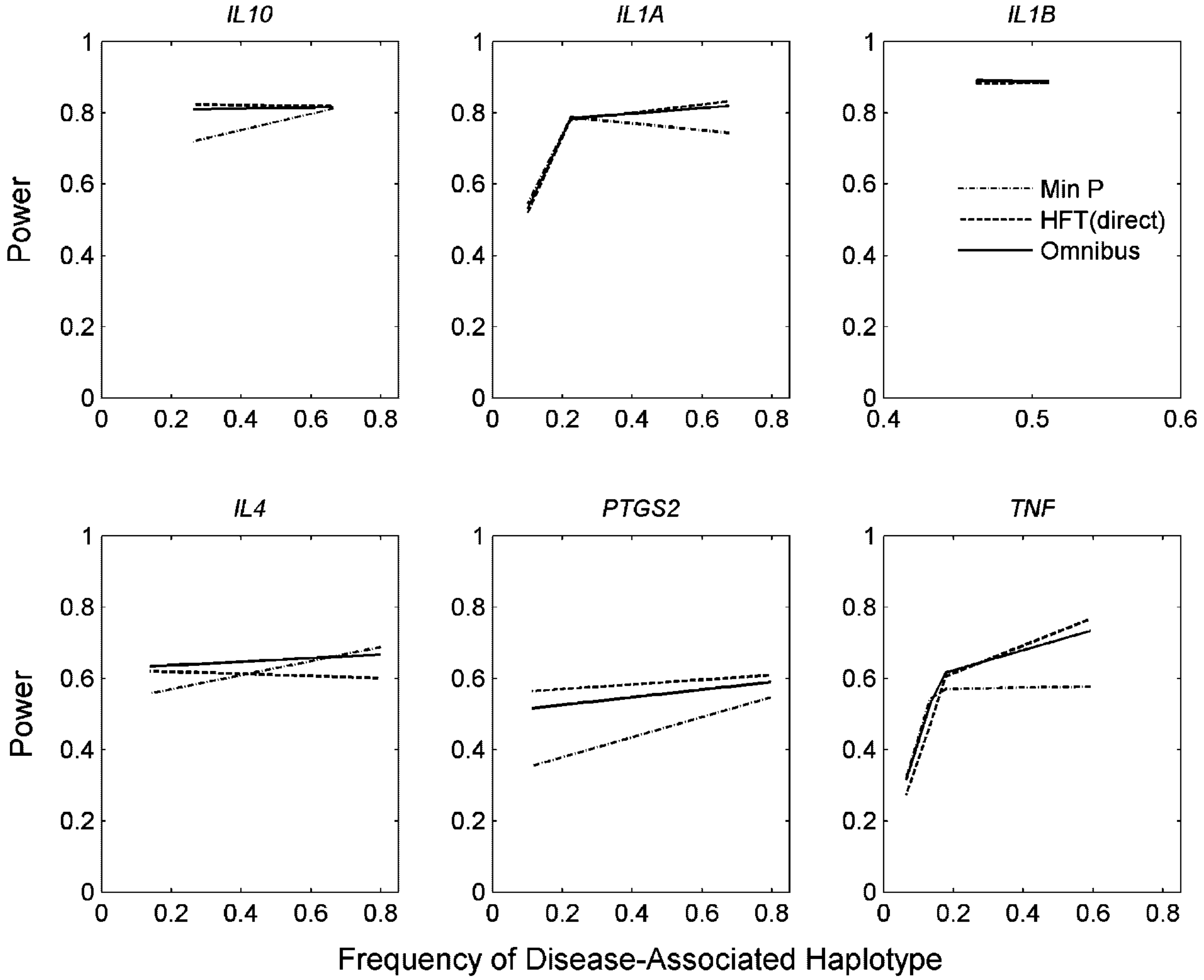

Fig. 5. Power curves of omnibus tests for the SBTCS panel when disease susceptibility is conferred by a haplotype with a co-dominant effect. The panel design is the same as in Figure 2. Power curves for min P test, directed HFT, and resampling-based omnibus test are shown.

TABLE III. SNP and haplotype analysis for SBTCS case-control study of bile duct cancer and variants in *PTGS2*

SNP Analysis				Haplotype Analysis		
SNP[a]	OR[b]	95%CI	Raw-p	Haplotype	Controls: frequency	Cancer cases: frequency
SNP1: G>C	1.48	0.79–2.80	0.2395	(00000)	0.7955	0.7182
SNP2: T>G	1.51	0.91–2.49	0.1214	(00010)	0.1113	0.1502
SNP3: T>C	1.56	0.90–2.68	0.1239	(01110)	0.0471	0.0709
SNP4: T>C	1.63	1.17–2.25	0.0041	(10000)	0.0356	0.0407
SNP5: C>T	1.03	0.23–4.67	0.9664	Other haplotypes	0.0096	0.0200

[a] SNP1: Ex3 −8 G>C; SNP 2: IVS5 −275 T>G; SNP3: IVS7 +111 T>C; SNP4: Ex10 +837 T>C; SNP5: Ex10 −90 C>T.
[b] Based on logistic regression with a co-dominant model.

studied genes and SNPs in the SNP500Cancer database.

A standalone computer program that implements these tests is available from B.E.C upon request. Limitations of our study should be noted. First, both **min P** and the resampling-based omnibus tests may have lengthy computation times, depending on the problem size. Second, the operating characteristics of the tests in more complicated models, such as models where association is due to multiple SNPs or haplotypes, or is induced by gene-environment interaction, have yet to be explored. Third, it is unclear by how much the power of the tests is diminished by the omission of SNPs within a gene that are not genotyped.

Nonetheless, for investigators with data in hand to address a priori hypotheses, the approaches we describe can be recommended. Although we did not present the details, each test can be extended to account for covariates, by using standard multiple logistic regression (**min P**) or by using for example the approach of Lake et al. [2003] (directed HFT). The enhanced tests may be particularly useful for genetic associations that appear to be of borderline significance using standard approaches. However, even if results are significant using standard tests, as illustrated by our analysis of data from the SBTCS, it may still be desirable to use the more powerful resampling-based approaches, for example, if the summary *p*-value for a single gene is to be adjusted for multiplicity over a panel of candidate genes. Finally, even in the largest meta-analysis of individual patient data or consortia study, the numbers quickly become sparse for subgroup analysis. Hence, the methods we propose may also be useful for subgroup analysis within large studies.

ACKNOWLEDGMENTS

The authors acknowledge Ron Hsu for assistance with simulation studies of haplotype analysis.

REFERENCES

Armitage P. 1955. Tests for linear trends in proportions and frequencies. Biometrics 11:375–386.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B (Methodol) 57:289–300.

De Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D. 2005. Efficiency and power in genetic association studies. Nat Genet 37:1217–1223.

Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B 39:1–38.

Emahazion T, Feuk L, Jobs M, Sawyer SL, Fredman D, St Clair D, Prince JA, Brookes AJ. 2001. SNP association studies in Alzheimer's disease highlight problems for complex disease analysis. Trends Genet 17:407–413.

Epstein MP, Satten GA. 2003. Inference on haplotype effects in case-control studies using unphased genotype data. Am J Hum Genet 73:1316–1329.

Excoffier L, Slatkin M. 1995. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12:921–927.

Fallin D, Cohen A, Essioux L, Chumakov I, Blumenfeld M, Cohen D, Schork NJ. 2001. Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. Genome Res 11:143–151.

Ge Y, Dudoit S, Speed TP. 2003. Resampling-based multiple testing for microarray data analysis. Test 12:1–77.

Hsing AW, Gao YT, Devesa SS, Jin F, Fraumeni JF Jr. 1998. Rising incidence of biliary tract cancers in Shanghai, China. Int J Cancer 75:368–370.

Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG. 2001. Replication validity of genetic association studies. Nat Genet 29:306–309.

Ioannidis JP, Rosenberg PS, Goedert JJ, O'Brien TR. 2002. Commentary: meta-analysis of individual participants' data in genetic epidemiology. Am J Epidemiol 156:204–210.

Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di GG, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, Twells RC, Payne F, Hughes W, Nutland S, Stevens H, Carr P, Tuomilehto-Wolf E, Tuomilehto J, Gough SC, Clayton DG, Todd JA. 2001. Haplotype tagging for the identification of common disease genes. Nat Genet 29:233–237.

Lake SL, Lyon H, Tantisira K, Silverman EK, Weiss ST, Laird NM, Schaid DJ. 2003. Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. Hum Hered 55:56–65.

Li SS, Khalid N, Carlson C, Zhao LP. 2003. Estimating haplotype frequencies and standard errors for multiple single nucleotide polymorphisms. Biostatistics 4:513–522.

Neale BM, Sham PC. 2004. The future of association studies: gene-based analysis and replication. Am J Hum Genet 75: 353–362.

Nicodemus KK, Liu W, Chase GA, Tsai YY, Fallin D. 2005. Comparison of type I error for multiple test corrections in large single-nucleotide polymorphism studies using principal components versus haplotype blocking algorithms. BMC Genet 6(suppl 1):S78.

Nyholt DR. 2004. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. Am J Hum Genet 74:765–769.

Prentice RL, Pyke R. 1979. Logistic disease incidence models and casecontrol studies. Biometrika 66:403–411.

R Development Core Team. 2004. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Rosenberg PS, Che A, Chen BE. 2006. Multiple hypothesis testing strategies for genetic case-control association studies. Stat Med, in press.

Sakoda LC, Gao YT, Chen BE, Chen J, Rosenberg PS, Rashid A, Deng J, Shen MC, Wang BS, Han TQ, Zhang BH, Cohen-Webb H, Yeager M, Welch R, Chanock S, Fraumeni JF Jr, Hsing AW. 2006. Prostaglandin-endoperoxide synthase 2 (PTGS2) gene polymorphisms and risk of biliary tract cancer and gallstones: a population-based study in Shanghai, China. Carcinogenesis.

SAS Institute Inc. 2002. SAS/Genetics User's Guide. Cary, NC: SAS Institute Inc.

Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. 2002. Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 70: 425–434.

Schork NJ. 2002. Power calculations for genetic association studies using estimated probability distributions. Am J Hum Genet 70: 1480–1489.

Schork NJ, Fallin D, Lanchbury JS. 2000. Single nucleotide polymorphisms and the future of genetic epidemiology. Clin Genet 58:250–264.

Simes RJ. 1986. An improved bonferroni procedure for multiple tests of significance. Biometrika 73:751–754.

Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978–989.

Westfall PH, Young SS. 1993. Resampling-Based Multiple Testing: Examples and Methods for *p*-Value Adjustment. New York: Wiley.

Westfall PH, Zaykin DV, Young SS. 2002. Multiple tests for genetic effects in association studies. Methods Mol Biol 184:143–168.

Zhao JH, Curtis D, Sham PC. 2000. Model-free analysis and permutation tests for allelic associations. Hum Hered 50: 133–139.

Zhao LP, Li SS, Khalid N. 2003. A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. Am J Hum Genet 72:1231–1250.

APPENDIX

DIRECTED HAPLOTYPE ANALYSIS

Suppose the haplotype frequencies are estimated using the EM algorithm [Excoffier and Slatkin, 1995]. The corresponding variance-covariance matrices for the estimated haplotype frequencies can be estimated by taking derivates of the observed log likelihood function at the solution. Let C be the total number of inferred haplotypes and

$$\hat{F} = (\hat{f}_1, \ldots, \hat{f}_C)$$

be the maximum likelihood estimate of the haplotype frequencies given the genotype data G. Without lost of generality, we suppose that $f_1 = (1 - \sum_{c=2}^{C} f_c) > 0$. For the jth genotype g_j with frequency p_j and observed counts n_j, its contribution to the log-likelihood function is given by $l_j = n_j \log(p_j)$. If g_j is the genotype consistent with homozygous haplotype pair $(h_1 h_1)$, its contribution to the information matrix derives from $l_j = 2n_j \log(1 - f_2 - \cdots - f_c)$ and is

$$\frac{\partial^2 l_j}{\partial f_s \partial f_t} = -2\frac{n_j}{f_1^2} \quad \text{for}\, s > 1,\ t > 1.$$

When g_j is consistent with haplotype pairs $(h_s h_s)$, for $s > 1$, we have

$$\frac{\partial^2 l_j}{\partial f_s^2} = -2\frac{n_j}{f_s^2} \quad \text{for all}\, s > 1.$$

If g_j is consistent with haplotype pair $(h_1 h_s)$, its contribution to the information matrix derives from $l_j = n_j \log\{\cdots + 2(1 - f_2 \cdots - f_s \cdots - f_t \cdots - f_C) f_s + \cdots\}$ and is

$$\frac{\partial^2 l_j}{\partial f_s^2} = -4\frac{n_j}{p_j^2}(f_1 - f_s)^2 - 4\frac{n_j}{p_j}$$

$$\frac{\partial^2 l_j}{\partial f_t^2} = -4\frac{n_j}{p_j^2} f_s^2 \quad s \neq t$$

$$\frac{\partial^2 l_j}{\partial f_s \partial f_t} = 4\frac{n_j}{p_j^2} f_s (f_1 - f_s) - 2\frac{n_j}{p_j}.$$

When g_j is consistent with haplotype pair $(h_s h_t)$, $s \neq t$, $s > 1$, $t > 1$, we have

$$\frac{\partial^2 l_j}{\partial f_s^2} = -4\frac{n_j}{p_j^2} f_t^2 \quad \text{and} \quad \frac{\partial^2 l_j}{\partial f_s \partial f_t} = -4\frac{n_j}{p_j^2} f_s f_t + 2\frac{n_j}{p_j}.$$

The corresponding variance-covariance matrix for the $(C-1)$ estimated haplotype frequencies $(\hat{f}_2, \ldots, \hat{f}_C)$ can be consistently estimated by

$$\hat{\sum}_{(C-1)} = \left[-\frac{\partial^2 l}{\partial f_s \partial f_t} \right]^{-1} \quad \text{for} \quad s > 1,\ t > 1.$$

To construct a directed HFT, let $\hat{F}_1$ and $\hat{F}_0$ be the estimated frequencies of the $P \leq C-1$ haplotypes of interest in the cases and the controls, respectively. The corresponding variance-covariance matrices can be estimated from the corresponding elements of $\hat{\sum}_{1,(C-1)}$ and $\hat{\sum}_{0,(C-1)}$ by $\hat{\sum}_{1,P}$ and $\hat{\sum}_{0,P}$. Denote the difference in frequencies between cases and controls by $\hat{\Delta} = \sqrt{n_1 n_0/(n_1 + n_0)}(F_1 - F_0)$, where n_1 and n_0 are number of cases and controls. Under H_0: $\Delta = 0$, the test statistic $\hat{\Delta}$ has asymptotic normal distribution with mean 0 and variance $\hat{\sum}_P = \hat{\sum}_{1,P} + \hat{\sum}_{0,P}$. Therefore, the Wald test statistic is given by

$$HFT_D = \hat{\Delta}' \hat{\sum_P}^{-1} \hat{\Delta} \xrightarrow{d} \chi_P^2.$$